# Implicit Optimization Bias of Next Token Prediction in Linear Models


## Abstract

Next-token prediction (NTP) has become the go-to training paradigm for modern language models, yet its optimization principles are not well-understood. To bridge this gap, we initiate a study of the structural properties of the solutions selected by gradient-based optimizers among the many possible minimizers of the NTP objective. By framing NTP as cross-entropy minimization across *distinct* contexts, each tied with a *sparse* conditional probability distribution across a finite vocabulary of tokens, we introduce "NTP-separability conditions" that enable reaching the entropy lower bound. With this setup, we then focus on linear models, for which we characterize the optimization bias of gradient descent. Extending previous research on implicit bias in one-hot classification to the NTP setting, highlights key differences and prompts further research into optimization and generalization of NTP.


## 1 Introduction

We initiate an investigation when implicit optimization biases in training language models under the NTP paradigm, particularly in overparameterized regimes where the empirical-loss reaches its lower bound and there is many possible minimizers. To formalize the NTP paradigm, consider autoregressive model $q_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ trained to predict the next-token on sequences of length $T$ using the cross-entropy (CE) loss:

$$\min_{\boldsymbol{\theta}} \hat{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{T}_n}\Big[\sum_{t\in[T]} -\log\left(q_{\boldsymbol{\theta}}(z_t\,|\,z_1,\ldots,z_{t-1})\right)\Big]. \quad (1)$$

Here, sequences $\boldsymbol{z} = (z_1,\ldots,z_T)$ consist of tokens $z_t$ from a finite vocabulary $\mathcal{V} = \{1,\ldots,V\}$ and $\hat{\mathbb{E}}$ is expectation over training set $\mathcal{T}_n$ of $n$ such sequences sampled from some underlying true distribution over sequences. Typically, the model $q_{\boldsymbol{\theta}}$ outputs probability of the next token computed via softmax applied on output logits, which are computed by projecting $d$-dimensional embeddings $h_{\boldsymbol{\theta}'}$ to the $V$-dimensional space with a trainable linear decoder $\boldsymbol{W} \in \mathbb{R}^{V\times d}$. Formally, [1] $q_{\boldsymbol{\theta}}(z_t\,|\,z_1,\ldots,z_{t-1}) = \mathbb{S}_{z_t}(\boldsymbol{W}h_{\boldsymbol{\theta}'}(z_1,\ldots,z_{t-1}))$. The CE loss is then minimized over $(\boldsymbol{W},\boldsymbol{\theta}')$ using gradient-based methods.

We pose the question: *Given training set $\mathcal{T}_n$, what are the structural properties of the weights $\boldsymbol{\theta}$ found by minimizing the NTP objective with gradient-based optimizers?* As in prior research in one-hot supervised classification [2] (e.g. (Zhang et al., 2017a; Belkin et al., 2018; Soudry et al., 2018; Ji and Telgarsky, 2018)), we specifically target this question in an *overparameterized* setting, where the NTP objective (1) may have an infinite number of solutions, representing an infinite number of models $\boldsymbol{\theta}$ that minimize the training loss. The central challenge is to discern the particular solution the optimizer is inherently biased towards. Since this 'bias' is not explicitly introduced through regularization but is instead ingrained in the training objective and algorithmic structure, it is termed 'implicit bias' (Neyshabur et al., 2014). The exploration of implicit bias has a long history in the traditional supervised one-hot classification (see *Related Work in Sec. 5*). In this traditional scenario, the training set comprises feature-label pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^p$ is a continuous feature, and $y$ represents its unique label. The optimization process minimizes the following training objective (over $\boldsymbol{W},\boldsymbol{\theta}'$): $\hat{\mathbb{E}}_{(\boldsymbol{x},y)}\left[-\log\left(\mathbb{S}_y(\boldsymbol{W}h_{\boldsymbol{\theta}'}(\boldsymbol{x}))\right)\right]$.

At first glance, excluding the sequential format of Eq. (1), the NTP training scenario might seem identical to traditional one-hot prediction: both aim to minimize the same CE loss across models that parameterize probabilities using the softmax of logits. Consider predicting the next token over fixed-length sequences, say sequences of length $t-1$, via optimizing: $\hat{\mathbb{E}}_{\boldsymbol{z}}\left[-\log\left(\mathbb{S}_{z_t}(\boldsymbol{W}h_{\boldsymbol{\theta}}(z_1,\ldots,z_{t-1}))\right)\right]$. The context here acts as the feature, and the next token as the label. Recent works (Liu et al., 2023; Malach, 2023) draw on such apparent similarities to the traditional one-hot classification paradigm to extrapolate known results from the latter to the NTP setting. However, this comparison overlooks a fundamental, yet critical difference in the nature of the training data that distinguishes these two paradigms (even when the sequential format of Eq. (1) is disregarded): In

[1] Throughout, $\boldsymbol{e}_v \in \mathbb{R}^V$ is the $v$-th standard basis vector, and $\mathbb{S}_z(\boldsymbol{u}) = \boldsymbol{e}_z^\top\mathbb{S}(\boldsymbol{u}) = e^{u_z}/\sum_{v\in V} e^{u_v}$ the $z$-th entry of softmax.

[2] In NTP, the ground-truth next token is inherently embedded within the underlying text, thus strictly speaking, it falls under the self-supervised learning paradigm (Radford et al., 2018). Yet, the utilization of the CE training objective bears striking resemblance to supervised training. We leverage this resemblance and regard NTP training as an instance of supervised learning, while also emphasizing how it differs from one-hot encoding supervision.

the traditional setting, each feature (e.g., image) is assigned a single label (e.g., image category). In contrast, in the NTP setting, contexts $z_1, \ldots, z_{t-1}$ of finite length sampled from finite vocabularies are naturally repeated in a (vast) training set, potentially multiple times, each time followed by *different* tokens $z_t$ (Shannon, 1951). Consequently, the NTP paradigm involves training over $m \leq n$ *distinct* (non-repetitive) contexts, each followed by a multitude of possible next tokens, appearing at varying frequencies. For instance, the context "`She is excellent at her role as a`" may be followed by next tokens such as "`doctor`," "`reviewer`," or "`mother`," each with different frequencies. Importantly, certain vocabulary tokens may *not* appear after a given context; e.g., in the above example, tokens like "`run`," "`and`," etc., will not follow.

**Model.** We study NTP training over a finite vocabulary employing the following model. Given a large training set of $n$ total sequences, we identify $m \leq n$ *distinct* contexts. Each distinct context $j \in [m]$ is linked to a $V$-dimensional empirical probability vector $\hat{\boldsymbol{p}}_j$, which encodes the frequency with which each vocabulary token follows the context throughout its occurrences in the training set. Crucially, the probability vectors $\hat{\boldsymbol{p}}_j$ are *sparse*, i.e., the support set $\mathcal{S}_j$ of $\hat{\boldsymbol{p}}_j$ satisfies $|\mathcal{S}_j| \ll |\mathcal{V}| = V$. In an extreme where $|\mathcal{S}_j| = 1, \forall j \in [m]$, the probability vector $\hat{\boldsymbol{p}}_j$ becomes one-hot, leading to a scenario reminiscent of the traditional classification setting described earlier. However, such an extreme is essentially improbable in practical language modeling (Shannon, 1951).

### 1.1 Contributions and Organization

**Formulation.** Recognizing the differences between NTP and one-hot classification, we study the question of implicit optimization bias within the NTP setting. To facilitate this, we utilize the model outlined in the previous paragraph and detailed in Sec. 2. For concreteness, our analysis adopts a 'top-down' approach, training only the decoding (also referred to as word-embedding) matrix $\boldsymbol{W} \in \mathbb{R}^{V \times d}$ while keeping context-embeddings fixed. This approach mirrors foundational studies on implicit optimization bias in one-hot classification (Soudry et al., 2018; Ji and Telgarsky, 2018), which first focused on linear models. It allows exploring the complexities of the NTP training objective, distinct from the embedding architecture[3], and while it renders the logits linear and the objective convex, it still poses a technical challenge in terms of determining parameter convergence (Soudry et al., 2018; Ji and Telgarsky, 2018; Ji et al., 2020).

**Conditions for reaching entropy.** In Sec. 3, we identify the necessary and sufficient conditions for the logits of the trained model to enable the CE loss to approach its lower bound, the empirical conditional entropy. We introduce two conditions: NTP$_{\mathcal{H}}$-compatibility and NTP-separability, which impose constraints on mutually orthogonal subspaces that are determined by the *sparsity patterns* of *distinct* contexts within the dataset. These conditions determine the necessary and sufficient overparameterization a model needs to achieve the empirical entropy lower bound during training.

**Margin in NTP setting and Implicit bias of GD.** Motivated by the NTP-separability condition, Sec. 4 introduces a margin concept for NTP, which extends the classical definition of margin used in one-hot supervised classification. We further establish the relevance of this new margin notion for GD optimization. Specifically, in the limit of iterations $k \to \infty$, we show the GD iterates grow undoubtedly in norm and converge to a finite $\boldsymbol{W}^\star$ within a data subspace $\mathcal{F}$, while simultaneously aligning with a NTP-margin maximizing direction $\boldsymbol{W}^{\mathrm{mm}}$ in the complementary subspace $\mathcal{F}^\perp$. The finite component $\boldsymbol{W}^\star \in \mathcal{F}$ solves a system of linear equations associated with NTP$_{\mathcal{H}}$-compatibility.

## 2 Setup

Let vocabulary $\mathcal{V} = [V] := \{1, \ldots, V\}$ represent a set of $V$ tokens (e.g. words) and $\boldsymbol{z}_{1:t} = (z_1, \ldots, z_t)$ denote sequence of $t$ tokens $z_t \in \mathcal{V}$. To simplify presentation, we focus on predicting the $T$-th token $z_T$ given contexts $\boldsymbol{z}_{<T} := \boldsymbol{z}_{1:T-1}$ of fixed length, and we further let $\boldsymbol{x} = \boldsymbol{z}_{<t}$ denote the context and $z$ denote the last token. See App. D for straightforward extension to the sequential format of Eq. (1). We assume access to a training set consisting of $n$ sequences $\mathcal{T}_n := \{(\boldsymbol{x}_i, z_i)\}_{i \in [n]}$, with $\boldsymbol{x}_i \in \mathcal{X} := \mathcal{V}^{T-1}$ and $z_i \in \mathcal{V}$. Let $h : \mathcal{X} \to \mathbb{R}^d$ an embedding map that maps contexts (i.e., sequences of $T-1$ tokens) to $d$-dimensional embeddings. The map $h$ is assumed fixed. The next-token is predicted via a linear model $f_{\boldsymbol{W}} : \mathcal{X} \to \mathbb{R}^V$ parameterized by decoding matrix $\boldsymbol{W} \in \mathbb{R}^{V \times d}$, such that $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W} h(\boldsymbol{x})$. When the model output passes through a softmax, it defines the model's probability mass function for the next-token prediction, given as $\hat{q}_{\boldsymbol{W}}(\cdot|\boldsymbol{x}) = \mathbb{S}(f_{\boldsymbol{W}}(\boldsymbol{x}))$, where $\mathbb{S}(\cdot) : \mathbb{R}^V \to \Delta^{V-1}$ is the softmax and $\Delta^{V-1}$ is the $V$-dimensional simplex. The decoder is trained by minimizing the empirical CE loss $\mathrm{CE}(\boldsymbol{W}) := \frac{1}{n} \sum_{i \in [n]} -\log\left(\hat{q}_{\boldsymbol{W}}(z_i|\boldsymbol{x}_i)\right)$.

Given dataset $\mathcal{T}_n$ we denote $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_m$ the $m \leq n$ *distinct* contexts among the (large number of) total $n$ contexts $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ within $\mathcal{T}_n$. Let $\hat{\pi}_j$ be the empirical probability of distinct context $\bar{\boldsymbol{x}}_j$. That is, $1 \leq n \cdot \hat{\pi}_j \leq n$ is the number of contexts $\boldsymbol{x}_i$ that equal $\bar{\boldsymbol{x}}_j$. Furthermore, for each distinct context $\bar{\boldsymbol{x}}_j, j \in [m]$ let $\hat{\boldsymbol{p}}_j \in \Delta^{V-1}$ denote the probability vector of conditional next-token distribution, i.e., $\hat{p}_{j,z} := \hat{p}\left(z|\bar{\boldsymbol{x}}_j\right), z \in \mathcal{V}, j \in [m]$. In other words, $n \cdot \hat{\pi}_j \cdot \hat{p}_{j,z}$ is the number of occurences of token $z$ as a follow-up to context $\bar{\boldsymbol{x}}_j$. Finally, we denote the support set and size of the support set of these conditional distributions as $\mathcal{S}_j := \{z \in \mathcal{V} \,|\, \hat{p}_{j,z} > 0\}$ and $S_j := |\mathcal{S}_j|$. Tokens $z \in \mathcal{S}_j$ and

[3] NTP is widely used across transformers (Radford et al., 2019), state-space models (Fu et al., 2022), LSTMs (Beck et al., 2024).

$v \notin \mathcal{S}_j$ are referred to as 'in-support' and 'out-of-support' respectively. With these, [4] we express the NTP loss as

$$\mathrm{CE}(\boldsymbol{W}) = -\sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z(\boldsymbol{W}\bar{\boldsymbol{h}}_j)\right), \quad (2)$$

where, we defined the shorthand $\bar{\boldsymbol{h}}_j = h(\bar{\boldsymbol{x}}_j)$. Similarly, we let $\boldsymbol{h}_i = h(\boldsymbol{x}_i), i \in [n]$. With some abuse of notation, we then obtain the following equivalent descriptions of the training set as $\mathcal{T}_m := \{(\bar{\boldsymbol{h}}_j, \hat{\pi}_j, \hat{p}_{j,z\in\mathcal{V}})\}_{j\in[m]}$ that emphasizes *distinct* contexts and their respective sparse next-token probability distributions.

The empirical $T$-gram entropy (referred to hereafter as entropy) of the dataset is (Shannon, 1948; 1951): $\mathcal{H}_T := \mathcal{H} := \hat{\mathbb{E}}_{(\boldsymbol{x},z)\sim\mathcal{T}_n}[-\log(\hat{p}(z|\boldsymbol{x}))] = -\sum_{j\in[m]}\sum_{z\in\mathcal{S}_j} \hat{\pi}_j \hat{p}_{j,z} \log(\hat{p}_{j,z})$. It lower bounds the CE loss since $\mathrm{CE}(\boldsymbol{W}) = \mathcal{H} + \mathrm{KL}(\hat{\boldsymbol{p}} \| \hat{\boldsymbol{q}}_{\boldsymbol{W}})$ and the KL divergence is nonnegative.

## 3 When is entropy reached?

Under what conditions on the training data can the CE loss reach its entropy lower-bound? By the entropy lower-bound, $\mathrm{CE}(\boldsymbol{W}) = \mathcal{H} \Leftrightarrow \mathrm{KL}(\hat{\boldsymbol{p}} \| \hat{\boldsymbol{q}}_{\boldsymbol{W}}) = 0$ iff for all $j \in [m]$ and all $z \in \mathcal{V}$: $\hat{q}_{\boldsymbol{W}}(z|\bar{\boldsymbol{x}}_j) = \hat{p}_{j,z}$. This leads to the following two sufficient and necessary conditions.

**Definition 3.1** ($\mathrm{NTP}_{\mathcal{H}}$-compatible)**.** $\mathcal{T}_m$ is $\mathrm{NTP}_{\mathcal{H}}$-compatible if $\exists \boldsymbol{W}^{\mathrm{p}} \in \mathbb{R}^{V\times d}$ satisfying $\forall j \in [m]$:

$$\forall z \neq z' \in \mathcal{S}_j : (\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \boldsymbol{W}^{\mathrm{p}} \bar{\boldsymbol{h}}_j = \log\left(\hat{p}_{j,z} / \hat{p}_{j,z'}\right). \quad (3)$$

**Definition 3.2** (NTP-separable)**.** $\mathcal{T}_m$ is NTP-separable if there exists $V \times d$ matrix $\boldsymbol{W}^{\mathrm{d}}$ satisfying the following:

$$\forall j \in [m], z \neq z' \in \mathcal{S}_j : (\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \boldsymbol{W}^{\mathrm{d}} \bar{\boldsymbol{h}}_j = 0 \quad (4a)$$

$$\forall j \in [m], v \notin \mathcal{S}_j : (\boldsymbol{e}_z - \boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{d}} \bar{\boldsymbol{h}}_j \geq 1. \quad (4b)$$

The eqns. in (3) constrain $\boldsymbol{W}^{\mathrm{p}}$ with respect to a subspace of $V\times d$ matrices that is defined in terms of context embeddings and their respective support sets:

$$\mathcal{F} = \mathrm{span}\left(\left\{(\boldsymbol{e}_z - \boldsymbol{e}_{z'})\bar{\boldsymbol{h}}_j^\top : z \neq z' \in \mathcal{S}_j, j \in [m]\right\}\right), \quad (5)$$

The subspace constraints in Eq. (4a) project $\boldsymbol{W}^{\mathrm{d}}$ onto the subspace $\mathcal{F}^\perp$, which is the orthogonal complement of the subspace $\mathcal{F}$ defined in (5). This leaves the softmax probabilities of possible next tokens (in set $\mathcal{S}_j$) intact, and fully determined by $\boldsymbol{W}^{\mathrm{p}}$ as per the $\mathrm{NTP}_{\mathcal{H}}$-compatibility condition. Formally, $\boldsymbol{W}^{\mathrm{p}} + \boldsymbol{W}^{\mathrm{d}}$ continues satisfying (3). Moving on the halfspace constraints in (4b), their impact on the softmax probabilities can be understood algebraically by considering $\boldsymbol{W}_\gamma := \gamma \boldsymbol{W}^{\mathrm{d}}$ and $v \notin \mathcal{S}_j$. We have: $\mathbb{S}_v(\boldsymbol{W}^\gamma \bar{\boldsymbol{h}}_j) = \left(\sum_{z\in\mathcal{S}_j} e^{\gamma(\boldsymbol{e}_z - \boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{d}} \bar{\boldsymbol{h}}_j} + \sum_{v'\notin\mathcal{S}_j} e^{\gamma(\boldsymbol{e}_v - \boldsymbol{e}_{v'})^\top \boldsymbol{W}^{\mathrm{d}} \bar{\boldsymbol{h}}_j}\right)^{-1} \leq e^{-\gamma}$, which approaches 0 as $\gamma \to \infty$.

[4] Please refer to Sec. E for list of notations.

**Proposition 3.3.** *Assume training data $\mathcal{T}_m$ is $\mathrm{NTP}_{\mathcal{H}}$-compatible and NTP-separable, with the respective matrices $\boldsymbol{W}^{\mathrm{p}}$ and $\boldsymbol{W}^{\mathrm{d}}$ satisfying conditions (3) and (4). While all finite $\boldsymbol{W}$ satisfy $\mathrm{CE}(\boldsymbol{W}) > \mathcal{H}$, it holds for $\boldsymbol{W}^\gamma = \boldsymbol{W}^{\mathrm{p}} + \gamma \cdot \boldsymbol{W}^{\mathrm{d}}$ that $\mathrm{CE}(\boldsymbol{W}^\gamma) \xrightarrow{\gamma\to+\infty} \mathcal{H}$.*

Thus, CE approaches its lower-bound in the limit of a *direction* $\overline{\boldsymbol{W}^{\mathrm{d}}} := \boldsymbol{W}^{\mathrm{d}}/\|\boldsymbol{W}^{\mathrm{d}}\|$ and *offset* $\boldsymbol{W}^{\mathrm{p}}$ satisfying the constraints of NTP-separability and NTP-compatibility, respectively. We remark that when $d > m$ (overparameterization) then the two constraints hold generically (see App. A.1).

## 4 Implicit Bias of GD

This section studies the implicit bias of GD. Denote the GD iterates at time $k$ by $\boldsymbol{W}_k = \boldsymbol{W}_{k-1} - \eta \nabla \mathrm{CE}(\boldsymbol{W}_{k-1})$ for arbitrary initial point $\boldsymbol{W}_0$ and constant step-size $\eta > 0$ small enough to guarantee descent. Under $\mathrm{NTP}_{\mathcal{H}}$-compatibility and NTP-separability, $\lim_{k\to\infty} \mathrm{CE}(\boldsymbol{W}_k) = \mathcal{H}$ and $\lim_{k\to\infty} \|\boldsymbol{W}_k\| = \infty$. This is intuitive because the CE loss is convex in $\boldsymbol{W}$ (thus, GD approaches the objective's infimum $\mathcal{H}$), and, in view of Proposition 3.3, the CE loss at all finite $\boldsymbol{W}$ is bounded away from $\mathcal{H}$. The relevant question then becomes that of determining the limit of the direction of the GD iterates. We need the following definitions: **(i)** $\boldsymbol{W}^\star \in \mathcal{F}$ is the unique solution of the $\mathrm{NTP}_{\mathcal{H}}$-compatibility equations on subspace $\mathcal{F}$. **(ii)** For NTP-separable training set $\mathcal{T}_m$, $\boldsymbol{W}^{\mathrm{mm}} \in \mathcal{F}_\perp$ is the unique solution to

$$\boldsymbol{W}^{\mathrm{mm}} := \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\| \quad \text{(NTP-SVM)}$$
$$\text{subj. to } \boldsymbol{W} \in \mathbb{R}^{V\times d} \text{ satisfying (4a) and (4b).}$$

This is a strongly convex quadratic program with $mV - \sum_{j\in[m]} S_j$ linear inequality and $\sum_{j\in[m]} S_j - m$ linear equality constraints. Its solution can be also defined as the classifier that maximizes margin between in and out-of -support tokens while being constrained on the orthogonal compelemnt $\mathcal{F}^\perp$: $\overline{\boldsymbol{W}^{\mathrm{mm}}} = \arg\max_{\|\boldsymbol{W}\|=1, \boldsymbol{W}\in\mathcal{F}^\perp} \min_{j\in[m], z\in\mathcal{S}_j, v\notin\mathcal{S}_j} (\boldsymbol{e}_z - \boldsymbol{e}_v)^\top \boldsymbol{W} \bar{\boldsymbol{h}}_j$.

**Theorem 4.1** (Implicit bias of GD)**.** *Assume $\mathrm{NTP}_{\mathcal{H}}$-compatible and NTP-separable training data $\mathcal{T}_m$. Then, it holds that $\lim_{k\to\infty} \left\langle \frac{\boldsymbol{W}_k}{\|\boldsymbol{W}_k\|}, \frac{\boldsymbol{W}^{\mathrm{mm}}}{\|\boldsymbol{W}^{\mathrm{mm}}\|} \right\rangle = 1$. Moreover, $\lim_{k\to\infty} \mathcal{P}_{\mathcal{F}}(\boldsymbol{W}_k) = \boldsymbol{W}^\star$.*

The theorem establishes that in the limit of iterations: $\boldsymbol{W}_k \approx \boldsymbol{W}^\star + \|\mathcal{P}_\perp(\boldsymbol{W}_k)\| \overline{\boldsymbol{W}^{\mathrm{mm}}}$, which is analogous to the result we obtained previously for the regularization path. The proof of the theorem's main ingredient (Lem. F.3 in the appendix) involves comparing the loss $\mathrm{CE}(\boldsymbol{W}_k)$ for large iterations $k$ to the loss evaluated at a "genie" point that is chosen so that: (i) On the subspace $\mathcal{F}$, it agrees with $\boldsymbol{W}_k$. This is because it is easy to show that $\mathcal{P}_{\mathcal{F}}(\boldsymbol{W}_k)$

converges to $\boldsymbol{W}^\star$ by standard gradient descent analysis for convex functions; (ii) On the orthogonal subspace $\mathcal{F}^\perp$, it follows the optimal (with respect to accelerating loss decrease) max-margin direction $\overline{\boldsymbol{W}^{\mathrm{mm}}} \in \mathcal{F}^\perp$. To establish the loss comparison, the ideas is to compare the values of the adjusted loss $\mathrm{CE}_\perp(\boldsymbol{W}) := \mathrm{CE}(\boldsymbol{W}) - \mathrm{CE}(\mathcal{P}_{\mathcal{F}}(\boldsymbol{W}))$.

See Sec. B for numerical examples verifying the theorem and visualizing the GD parameters as word-embeddings (e.g., Fig. 1, 2).

## 5 Related work

We build on the literature on implicit optimization bias of CE loss in one-hot supervised classification. (Soudry et al., 2018) show that for linear models and linearly-separable data, GD converges in direction to the max-margin classifier. This result strengthens (Rosset et al., 2003) that showed the regularization path of CE minimization converges to the same limit. Closer to us, (Ji and Telgarsky, 2018; Ji et al., 2020) extend the analysis to encompass general binary data as follows: the data are linearly separable only on a certain subspace, and they show that GD converges, in direction, towards the max-margin classifier confined within that subspace. On the orthogonal subspace, it converges to a finite point. While operationally similar, our finding in Thm. 4.1 *cannot* be directly derived from theirs since our setting is neither binary nor one-hot. Nevertheless, our proofs extend the foundational work of (Rosset et al., 2003; Ji and Telgarsky, 2018; Ji et al., 2020), akin to numerous other studies that explore extensions to nonlinear architectures, (Lyu and Li, 2020; Ji and Telgarsky, 2020; Gunasekar et al., 2018a;b; Tarzanagh et al., 2023b) and to stochastic and adaptive algorithms, e.g. (Nacson et al., 2019; Pesme et al., 2021; Damian et al., 2021; Sun et al., 2022). The implicit bias viewpoint has also created opportunities to study generalization in overparameterized settings. (Hastie et al., 2019; Bartlett et al., 2019; Montanari et al., 2019) build a two-stage approach initially leveraging implicit bias to simplify the complexities of optimization before addressing generalization. This narrows the generalization question to the properties of the corresponding max-margin classifier , e.g. (Muthukumar et al., 2020; Cao et al., 2021; Koehler et al., 2021; Donhauser et al., 2022). The same strategy has also been adopted to study model robustness to adversarial perturbations (Javanmard and Soltanolkotabi, 2022; Taheri et al., 2023; Chen et al., 2023), out-of-distribution data (Tripuraneni et al., 2021), and imbalances (Sagawa et al., 2020; Chatterji et al., 2021; Kini et al., 2021). Our results motivate such extensions in the richer NTP setting.

Finally, while fundamentally different in nature, the form of our convergence results echoes a recent conjecture by (Tarzanagh et al., 2023a) regarding implicit optimization bias in transformers. Unlike their conjecture, which focuses on binary classification, our results are rigorously proven and apply to the NTP setting. Further detailed discussion on related follow-up work is deferred to Appendix C.

## 6 Future work

As the first study of implicit biases in NTP training, we use several assumptions essential for establishing an initial foundational understanding. The framework allows for various exciting promising research directions, some of which we outline below. Within the linear setting and GD:

• **NTP-separability thresholds:** Identifying exact thresholds for NTP-separability under distributional assumptions, akin to previous work on one-hot separability (Remark A.2).

• **Generalization:** Studying generalization in NTP settings by examining statistical properties of the NTP-SVM solution. Past research has successfully undertaken similar investigations for one-hot classification (see Sec. 5). While we acknowledge the importance of addressing specific challenges inherent to NTP —such as determining an appropriate measure of generalization, or establishing suitable statistical models for context-embeddings that respect the discrete nature of the underlying token subsequences—we believe this direction holds promise for further exploration.

Also, towards relaxing the linearity assumption:

• **Architecture-specific embeddings:** In a bottom-up approach that considers architecture-specific embeddings, the initial step involves modeling embeddings produced by, for instance, a shallow transformer, and examining the effects of regularization biases on the training of both the transformer and the decoder weights. An important nuance in this approach is the need to restrict to shallow transformers in a manner that still enables the NTP loss to attain the entropy lower bound. This may necessitate limiting the training data distribution, e.g. (Makkuva et al., 2024).

• **Memory capacity in NTP settings:** Without restricting the data beyond the discrete nature of tokens from a finite vocabulary, there is a strong case for investigating memory capacity of transformers in NTP, where recent studies on transformer memory capacity (Kajitsuka and Sato, 2024; Kim et al., 2023) do *not* apply.

• **Unconstrained features:** Extending the top-down approach, one could consider freely optimizing context embeddings together with decoder vectors (also known as word embeddings). The resulting log-bilinear model, reminiscent of wor2vec models (Pennington et al., 2014; Mikolov et al., 2013), extends the unconstrained features model, which has recently been employed to investigate neural collapse geometry in one-hot classification settings. This idea offers a promising avenue for uncovering structures in the geometries of context and word embeddings when learned jointly, potentially revealing new insights into the capabilities of sufficiently expressive language models.

• **Other optimizers:** Exploring the NTP implicit bias of adaptive algorithms, such as Adam.

## References


Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

Burak Çakmak, Yue M Lu, and Manfred Opper. A convergence analysis of approximate message passing with non-separable functions and applications to multi-class classification. *arXiv preprint arXiv:2402.08676*, 2024.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *The Journal of Machine Learning Research*, 22(1):7135–7182, 2021.

Jinghui Chen, Yuan Cao, and Quanquan Gu. Benign overfitting in adversarially robust linear classification. In *Uncertainty in Artificial Intelligence*, pages 313–323. PMLR, 2023.

Sitan Chen and Yuanzhi Li. Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*, 2024.

Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52, 2022.

Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning curves for the multi-class teacher–student perceptron. *Machine Learning: Science and Technology*, 4 (1):015019, 2023.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, pages 326–334, 1965.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning*, pages 5397–5428. PMLR, 2022.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*, 2021.

Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31:9461–9471, 2018b.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? 2024.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. 2023.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34: 18970–18983, 2021.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34: 20657–20668, 2021.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning, 2023.

Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pages 685–693. PMLR, 2024.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive power of self-attention matrices, 2021.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR, 2023.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024.

Eran Malach. Auto-regressive next-token predictors are universal learners. *arXiv preprint arXiv:2309.06979*, 2023.

Francesca Mignacco, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544*, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are

unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Saharon Rosset, Ji Zhu, and Trevor J. Hastie. Margin maximizing loss functions. In *NIPS*, 2003.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. *International Conference on Machine Learning*, 2022.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 976–980. IEEE, 2018.

Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

Viktoriia Sharmanska, Daniel Hernández-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2202, 2016.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary linear classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Kai Tan and Pierre C Bellec. Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. *arXiv preprint arXiv:2305.17825*, 2023.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023a.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism, 2023b.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer, 2023a.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34:13883–13897, 2021.

R. Vershynin. Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal. umich. edu/romanv/papers/GFA-book/GFA-book. pdf*, 2011.

Johannes von Oswald, Eyvind Niklasson, E. Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *ArXiv*, abs/2212.07677, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017a.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context, 2023.

# A Additional Results

## A.1 On the role of overparameterization

We show that overparameterization provides a sufficient condition for the solvability of (3) and (4).

Start with the halfspace constraints in (3) for NTP$_{\mathcal{H}}$-compatibility. These can be compactly expressed as $\boldsymbol{E}_{j,z_j}\boldsymbol{W}^{\mathrm{p}}\bar{\boldsymbol{h}}_j = \boldsymbol{a}_{j,z}$, where $\boldsymbol{E}_{j,z_j} \in \mathbb{R}^{(S_j-1)\times V}$ has rows $\boldsymbol{e}_{z_j} - \boldsymbol{e}'_z$ and $\boldsymbol{a}_{j,z_j} \in \mathbb{R}^{(S_j-1)}$ has entries $\log\left(\hat{p}_{j,z_j}/\hat{p}_{j,z'}\right)$ for some anchor $z_j \in \mathcal{S}_j$. Now, since the rows of $\boldsymbol{E}_{j,z_j}$ are linearly independent, the question becomes equivalently that of determining when $\boldsymbol{W}^{\mathrm{p}}\left[\bar{\boldsymbol{h}}_1,\dots,\bar{\boldsymbol{h}}_m\right] = \left[\boldsymbol{E}^{\dagger}_{1,z_1}\boldsymbol{a}_1,\dots,\boldsymbol{E}^{\dagger}_{m,z_m}\boldsymbol{a}_m\right]$ has a solution. This is always the case when $d > m$ and the $d \times m$ embedding matrix $\bar{\boldsymbol{H}} = \left[\bar{\boldsymbol{h}}_1,\dots,\bar{\boldsymbol{h}}_m\right]$ is full rank ($m$). Then, there exists $\boldsymbol{W}^{\mathrm{p}}$ such that condition (3) holds. In fact, $\bar{\boldsymbol{H}}^{\top}$ has a nullspace, implying the existence of an infinite number of solutions to (3). These solutions take the form $\boldsymbol{W}^{\mathrm{p}} = \boldsymbol{W}^{\star} + \boldsymbol{W}^{\mathrm{p}}_{\perp}$, where $\boldsymbol{W}^{\star} \in \mathcal{F}$ is the unique solution onto the subspace, and $\boldsymbol{W}^{\mathrm{p}}_{\perp} \in \mathcal{F}^{\perp}$. In contrast to (3), the constraints in (4) involve linear inequalities. However, a sufficient proxy for feasibility in this case is that the corresponding system of equations (instead of inequalities) has a solution. By following the exact same argument as before, we arrive at the same sufficient conditions for the existence of a solution $\boldsymbol{W}^{\mathrm{d}}$. We summarize these findings.

**Lemma A.1** (Overparameterization implies NTP-separability)**.** *Assume overparameterization $d > m$ and full-rank embedding matrix $\bar{\boldsymbol{H}} \in \mathbb{R}^{d\times m}$. Then, there exists an infinite number of solutions $\boldsymbol{W}^{\mathrm{p}}$ and $\boldsymbol{W}^{\mathrm{d}}$ that satisfy conditions (3) and (4), respectively.*

Thus, $d > m$, [5] which also generically favors full-rankness of the embedding matrix (Vershynin, 2011), implies both NTP$_{\mathcal{H}}$-compatibility and NTP-separability. Combined with Prop. 3.3, it also implies that there are infinitely many possible directions $\boldsymbol{W}^{\mathrm{d}}$ along which the NTP loss approaches $\mathcal{H}$, motivating the implicit-bias question: For a specific iterative algorithm aimed at minimizing the NTP loss, which direction does it prefer? We will address this question in the remainder of the paper.

*Remark* A.2. In the trivial case where $S_j = 1, \forall j \in [m]$ (one-hot classification), the entropy lower bound is zero and is attained iff the data is linearly separable. Indeed, $\mathcal{F}$ reduces to the empty set, and NTP-separability simplifies to traditional multiclass separability. For binary classification, (Cover, 1965) showed that $d/m > 1/2$ is sufficient and necessary for data in general position to be linearly separable. More recently, several works have extended this analysis to structured (random) data, including (Candès and Sur, 2018; Salehi et al., 2018; Montanari et al., 2019; Mignacco et al., 2020). The exact threshold in corresponding mutliclass settings is more intricate, but (Cornacchia et al., 2023; Tan and Bellec, 2023; Çakmak et al., 2024) have made progress in this direction. An interesting question is determining exact thresholds for NTP-separability, which would improve upon the sufficient condition of Lemma A.1.

## A.2 Regularization path

This section investigates the implicit bias of NTP by examining the minimization of CE loss through iterates defined as follows for an increasing sequence of positive regularization parameters $B$:

$$\widehat{\boldsymbol{W}}_B := \arg\min_{\|\boldsymbol{W}\|\le B} \mathrm{CE}(\boldsymbol{W})\,. \tag{6}$$

This involves minimizing a strictly convex function in a bounded domain; thus, $\widehat{\boldsymbol{W}}_B$ is unique. This section's main result characterizes the limit of $\widehat{\boldsymbol{W}}_B$ as $B \to \infty$ under NTP-separability/compatibility. Before that, we first define the next-token prediction support-vector machines (SVM) problem.

**Theorem A.3** (Implicit bias of the regularization-path)**.** *Assume training data $\mathcal{T}_m$ is NTP$_{\mathcal{H}}$-compatible and NTP-separable. Let $\widehat{\boldsymbol{W}}_B$ be defined as in (6). Then, it holds that* $\lim_{B\to\infty}\left\langle \frac{\widehat{\boldsymbol{W}}_B}{\|\widehat{\boldsymbol{W}}_B\|}, \frac{\boldsymbol{W}^{\mathrm{mm}}}{\|\boldsymbol{W}^{\mathrm{mm}}\|}\right\rangle = 1\,.$

# B Experiments

All experiments were conducted on a MacBook Pro equipped with a 2.3 GHz Quad-Core Intel Core i7 processor and 32 GB of memory. The experiments are of relatively small scale and were implemented in Matlab. The code is straightforward to reproduce, following the detailed specifications provided in the subsequent sections.

[5] The necessity for such large $d$ can be mitigated through the utilization of non-linear architectures (such as an MLP decoder), in which the total number of parameters can be increased by augmenting the width or depth, rather than directly modifying the embedding dimension $d$ as in linear models.

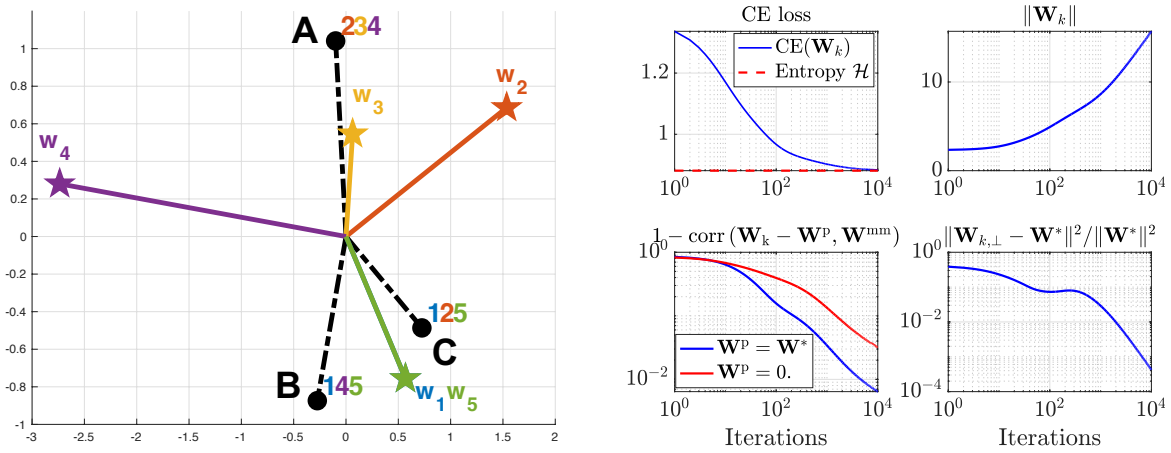


Figure 1: Vis. of NTP implicit optimization bias in a setting with $m = 3$ *distinct* contexts, embedding dimension $d = 2$, vocabulary of $|\mathcal{V}| = 5$ words and support sets of length $|\mathcal{S}_j| = 3, j \in [3]$. ***Left:*** Vis. of *context embeddings* $\bar{\boldsymbol{h}}_j$ in circle black markers (marked as A,B,C) and of their associated support sets $\mathcal{S}_j$ (colored text below each marker). Colored vectors (star markers) represent max-NTP-margin vectors $\boldsymbol{w}_v^\top := \boldsymbol{e}_v^\top \boldsymbol{W}^{\mathrm{mm}}, v \in [5]$ found by GD. Interpreting decoder vectors as *word embeddings* leads to intuitive findings on their geometry learned by NTP training. E.g., word embedding $\boldsymbol{w}_3$ (almost) aligns with context-embedding $A$ and the normal hyperplane it defines separates $A$ from $B$ and $C$, since word 3 only appears after context $A$. The rest of the words follow two contexts each and their word-representation naturally belongs to the cone defined by the embeddings of those respective contexts. The wider the cone, the larger the magnitude of the word embedding to compensate for the large angle between context-representations that share the same next-word. Note that geometry of depicted word embeddings only depends on support sets, but the conditional probabilities define another set of word representations on an orthogonal (matrix) subspace; see text for details and vis. ***Right:*** Upper/lower graphs confirm the predictions of Theorem 4.1.

### B.1 2D visualization: Interpretation of word-embeddings

Figure 1 illustrates a toy 2d example where the embeddings and the hyperplanes defined by each row of $\boldsymbol{W}^{\mathrm{mm}}$ can be visualized. We used $d = 2, m = 3, V = 5$ and $S_1 = S_2 = S_3 = 3$. The right subfigure shows results of GD training with respect to training loss, norm growth, alignment with $\boldsymbol{W}^{\mathrm{mm}}$, and convergence to $\boldsymbol{W}^\star$ on $\mathcal{F}$. See App. B for further implementation details and additional experiments. The left subfigure illustrates: (i) In black markers, the context-embedding geometry along with the associated support sets for each context A, B, and C. (ii) In colored markers, the geometry of word-embeddings, that is the max-NTP-margin vectors $(\boldsymbol{W}^{\mathrm{mm}})^\top \boldsymbol{e}_v, v \in [5]$, to which GD directionally converges. See caption for interpretation.

Additionally, Figure 2 shows the matrix of conditional probabilities and visualizes next to each other (i) the rows of the directional component $\boldsymbol{W}^{\mathrm{mm}}$ (Middle) and (ii) those of the finite component $\boldsymbol{W}^\star$ (Right). Interpreting the $V \times d$ decoder matrix as the matrix of learned word embeddings, this provides a visualization of their geometry. As per our results, the two word-embedding matrices $\boldsymbol{W}^\star$ and $\boldsymbol{W}^{\mathrm{mm}}$ lie on orthogonal subspaces. The geometry of the first depends on the probabilities of in-support tokens, while that of the second depends only on the support set of these probabilities. See also caption of Fig. 2.

### B.2 Overparameterized setting

We examine the implicit bias of GD on NTP training with overparameterization on synthetic data generated as follows. We construct dataset with $n = 5000$ sequences involving $m = 50$ distinct contexts. Each distinct context gets mapped to a randomly generated embedding of dimension $d = 60 > m$. We set vocabulary size $V = 10$ and each context $j \in [m]$ is followed by $S_j = 6, \forall j \in [m]$ possible next-tokens. The support sets $\mathcal{S}_j \subset \mathcal{V}$ and the probabilities $\hat{p}_{j,z}, z \in \mathcal{S}_j$ are chosen randomly; see Fig. 3 for representative examples from the training dataset. For a fixed realization of the dataset (for which $\mathcal{H} \approx 1.445$nats), we run GD, normalized GD (NGD), and Adam from random LeCun initialization. For GD, we use learning rate $\eta = 0.5$ and for NGD and Adam $\eta = 0.01$. For Adam, we also set $\beta_1 = 0.9, \beta_2 = 0.99$. We run all algorithms for $1e4$ iterations. For each case, we plot the following as a function of iterations:

1. Upper Left: CE loss versus entropy lower bound

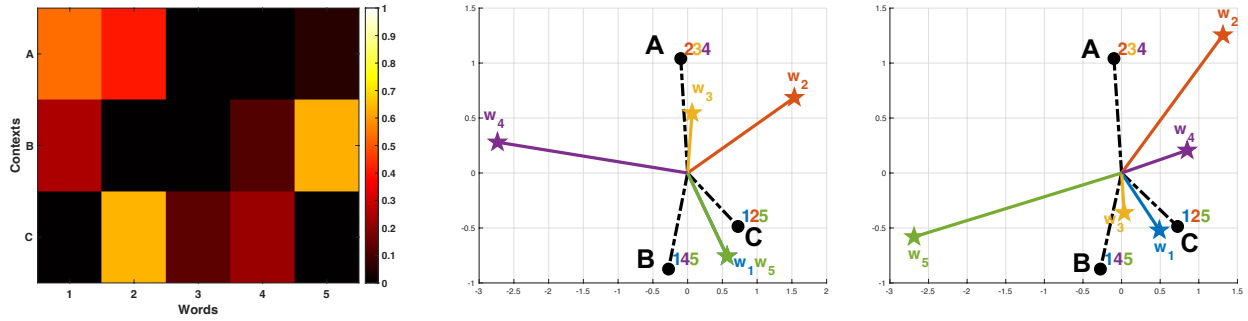


Figure 2: Same setup as Fig. 1. ***Left:*** Matrix $\boldsymbol{P}$ of conditional probabilities of words (cols.) per context (rows). Each row corresponds to the conditional probability vectors $\boldsymbol{p}_j, j \in [m]$. Black entries correspond to off-support words. ***Middle:*** Shown as $\boldsymbol{w}_z, z \in [5]$, the rows of the NTP-SVM solution $\boldsymbol{W}^{\mathrm{mm}}$ to which GD directionally converges. ***Right:*** Shown as $\boldsymbol{w}_z, z \in [5]$, the rows of the finite parameter $\boldsymbol{W}^\star$ to which GD iterates projected on $\mathcal{F}$ converge to. The geometry of $\boldsymbol{W}^{\mathrm{mm}}$ depends only on the support-set of $\boldsymbol{P}$. On the other hand, the geometry of $\boldsymbol{W}^\star$ depends on the entries of $\boldsymbol{P}$ for in-support tokens/words. As seen from visualization of $\boldsymbol{P}$, the words 1 and 5 have the same support pattern (i.e., both follow the same contexts $A$ and $B$). Thus, $\boldsymbol{w}_1 = \boldsymbol{w}_5$ in the Middle plot. However, on the subspace $\mathcal{F}$ corresponding to the Right plot, $\boldsymbol{w}_1 \neq \boldsymbol{w}_5$, which allows matching the different conditional probabilities with which each follows contexts $A$ and $B$.

2. Upper Right: parameter norm growth
3. Lower Left: correlation of $\boldsymbol{W}^{\mathrm{mm}}$ with iterates $\boldsymbol{W}_k$ and of "corrected" iterates $\boldsymbol{W}_k - \boldsymbol{W}^\star$ after substracting the component on $\mathcal{H}$
4. Lower Right: convergence of the subspace component $\boldsymbol{W}_{k,\mathcal{F}} = \mathcal{P}_{\mathcal{F}}(\boldsymbol{W}_k)$.

Fig. 4 shows an instance of these. As predicted by our analysis, in this overparameterized setting: CE loss converges to its lower-bound, parameter norm increases, iterates align in direction with $\boldsymbol{W}^{\mathrm{mm}}$, and the subspace component converges to $\boldsymbol{W}^\star$.

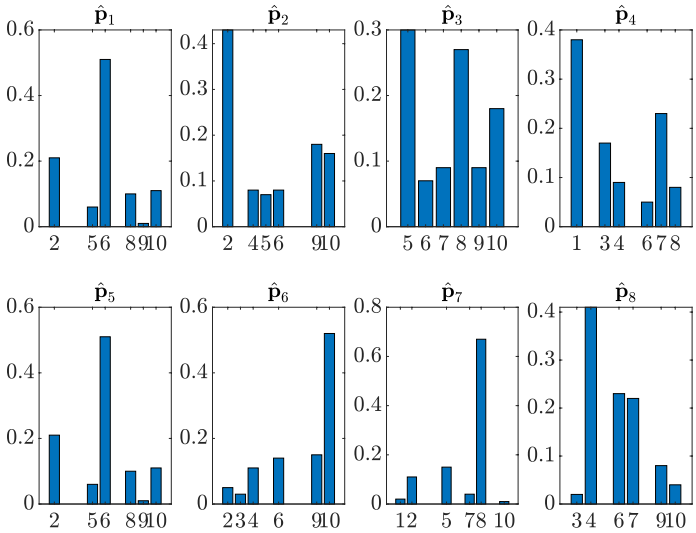


Figure 3: Eight randomly picked contexts with their associated next-token empirical conditional probabilities $\hat{\boldsymbol{p}}_j$. The indices shown on the x-axis define the support set $\mathcal{S}_j$ of each context.

Figure 5 illustrates the same plots, but this time for training over the same dataset with NGD and Adam. We observe same

implicit bias, but faster convergence. For NGD, this is consistent with analogous findings (rigorous in that case) for one-hot classification (Nacson et al., 2019; Ji and Telgarsky, 2021).

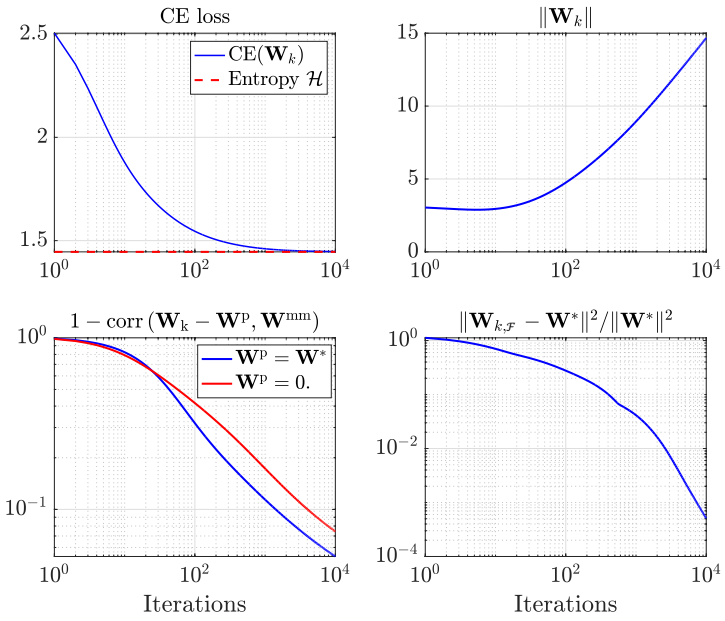


Figure 4: Experimental illustration of the implicit bias of GD in NTP over synthetic data with overparameterization. See App. B for detailed description of the experimental setting.

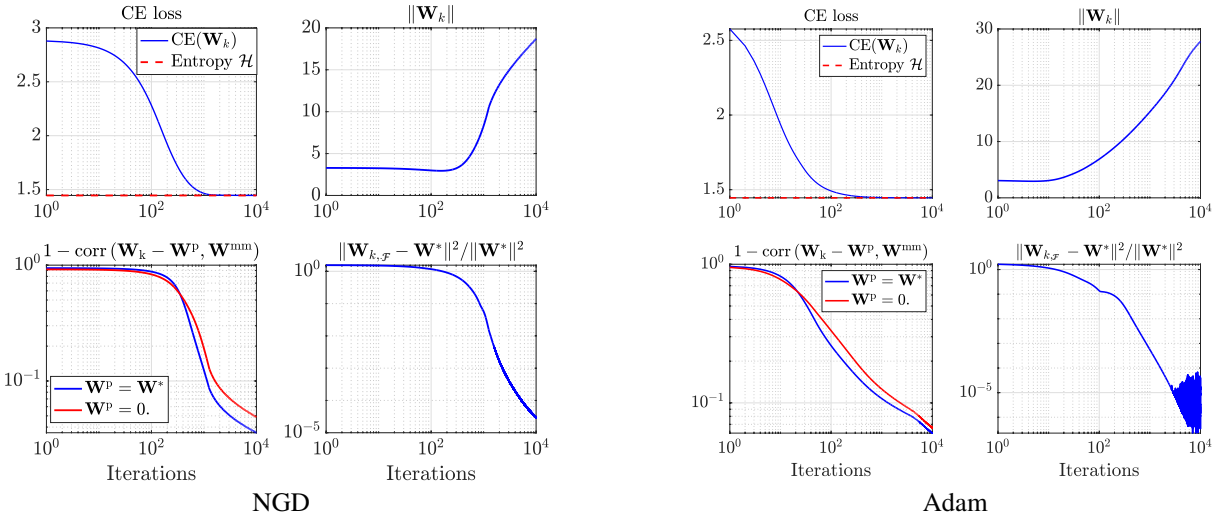


Figure 5: Implicit bias of *normalized* GD (Left) and of Adam (Right) in NTP over synthetic data with overparameterization. Both exhibit the same implicit bias, but converge faster than GD, with Adam being slightly faster than NGD.

## C Additional related work

**Implicit bias in transformers.** As already mentioned in Sec. 5, our work is closely related to (Tarzanagh et al., 2023a), where the authors investigate the implicit bias of self-attention in transformers. The insight put forth in the prequel (Tarzanagh et al., 2023b) is that softmax attention induces implicit-bias behaviors that bear similarities to vanilla implicit bias of one-hot prediction. Concretely, (Tarzanagh et al., 2023a) studies GD optimization of one-layer self-attention with

fixed decoder and *one-hot binary* classification. They show that, in the limit, GD finds attention weights that converge in direction to the solution of an SVM problem that separates optimal tokens from non-optimal ones. Their non-convex setting introduces locally optimal SVM directions to which GD may converge depending on initialization. Different to them, the NTP setting that we study involves predictions over multiple categories and is *not* one-hot. Also, while they fix the decoder, here, we fix the embeddings. In these respects their results are rather different. More similarities arise when (Tarzanagh et al., 2023a) replace the linear decoder with a MLP, which they note can induce multiple optimal tokens per sequence. This leads them to formulate a more general token-separating SVM program, which similar to ours confines the separation on a certain data subspace. However, the operational nature of the programs remains different as theirs optimizes attention weights and separates tokens within a sequence, while ours optimizes decoder weights and separates context embeddings based on their respective support sets. More importantly, while (Tarzanagh et al., 2023a) only conjectures the convergence of GD to their general SVM program, we leverage convexity in our setting to prove an analogous statement rigorously. Eventually, as we move lower in our top-down approach and consider architecture-specific embeddings generated by attention, we anticipate to see integration of our ideas with theirs.

Beyond (Tarzanagh et al., 2023a), there is growing recent research investigating optimization and generalization principles of transformers, e.g., (Sahiner et al., 2022; Edelman et al., 2021; Likhosherstov et al., 2021; von Oswald et al., 2022; Zhang et al., 2023; Akyürek et al., 2023; Li et al., 2023; Tarzanagh et al., 2023b;a; Tian et al., 2023a; Chen and Li, 2024). These efforts predominantly employ a 'bottom-up' approach that involves isolating shallow transformers, often with simplifications such as removing MLPs, utilizing single heads instead of multiple, and fixing certain parts while training only a subset of trainable parameters. Most of these studies have focused on classical one-hot supervised settings, and only a handful (e.g., (Tian et al., 2023a;b)) have seeked extending these 'bottom-up' analyses to NTP settings. Yet, their primary emphasis remains on uncovering the role of attention and how attention weights evolve during training. Instead, our approach uniquely emphasizes the NTP training paradigm itself, shifting the focus from the intricacies of specific transformer architectures.

Upon completing this paper, we became aware of independent contemporaneous research by Li et al. (Li et al., 2024) that also examines the implicit bias of self-attention with a fixed linear decoder in next-token prediction scenarios. Unlike our study which utilizes the widely adopted CE loss, their approach is based on log-loss, which renders the training loss convex, a similarity shared with our model despite the inclusion of self-attention. Both our results and those of Li et al. substantiate the conjecture posited by Tarzanagh and colleagues (Tarzanagh et al., 2023a), albeit in very distinct settings. Notably, contrary to both (Tarzanagh et al., 2023b) and (Li et al., 2024), we unveil the optimization intricacies of the NTP paradigm, even within the simplest linear settings.

**Classification with soft labels.** Unlike one-hot classification, soft-label classification associates each example with a probability vector, where each entry represents the likelihood of a corresponding label characterizing the example. Although arguably less prevalent than one-hot (or hard-label) classification, soft-label classification arises in various contexts, including modeling human confusion during crowd-sourcing (Peterson et al., 2019; Sharmanska et al., 2016; Collins et al., 2022), knowledge distillation (Hinton et al., 2015), label smoothing (Szegedy et al., 2016), and mixup (Zhang et al., 2017b). Our model of last-token prediction also falls within this setting. Specifically, our approach is most closely related to soft-labels generated by averaging annotators' hard labels (Peterson et al., 2019), rather than following the winner-takes-all rule to assign labels. (Peterson et al., 2019) and follow-up work have provided empirical evidence that using probabilistic soft labels generated from crowd annotations for training leads to improved performance in terms of model generalization, calibration, and robustness to out-of-distribution data. To the best of our knowledge, no prior work has investigated the implicit bias of gradient descent in this or other soft-label classification settings; thus, our results are of direct relevance to these contexts as well.

## D Autoregressive setting

For concreteness and simplified notation, in the paper's main body we focus on NTP over sequences of fixed length. We show here that this encompasses the autoregressive (i.e., sequential) setting with minimal changes. This also emphasizes the role played in our results by the sequence length.

As pointed in (1), the full autoregressive NTP objective averages $T$ individual losses (without loss of generality assume sequences of equal maximum length $T$). In order to make our analysis applicable, we first need to express (1) in terms of *unique* contexts. Mirroring the notations in Sec. 2, define the following for $t \in [T-1]$:

- $m_t, t \in [T-1]$ is the number of *distinct* contexts of size $t$. Note that $m_1 \geq m_2 \geq \cdots \geq m_{T-1}$.

- $m = \sum_{t=1}^{T-1} m_t$ is the total number of distinct contexts in the dataset
- $\bar{\boldsymbol{h}}_{t,j} := \boldsymbol{h}_{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}_{j,t}), t \in [T-1], j \in [m_t]$ is the embedding of the $j$-th (among all $t$-long contexts) distinct context $\bar{\boldsymbol{x}}_{j,t}$.
- $\hat{\pi}_{j,t}$ is the empirical probability of $\bar{\boldsymbol{x}}_{j,t}$.
- $\hat{p}_{j,t,z}$ is the empirical probability that context $\bar{\boldsymbol{x}}_{j,t}$ is followed by token $z \in \mathcal{V}$.
- $\mathcal{S}_{j,t}$ is the support set of the next-token distribution of context $\bar{\boldsymbol{x}}_{j,t}$.

With this notation, the NTP objective becomes

$$\mathrm{CE} = -\sum_{t\in[T-1]} \sum_{j\in[m_t]} \hat{\pi}_{t,j} \sum_{z\in\mathcal{S}_{j,t}} \hat{p}_{t,j,z} \log\left(\mathbb{S}_z(\boldsymbol{W}\bar{\boldsymbol{h}}_{t,j})\right) .$$

To continue enumerate the multi-set $\mathcal{I} := \{i = (j,t) \,|\, t \in [T-1], j \in [m_t]\}$. We may then rewrite the above as

$$\mathrm{CE} = -\sum_{i\in\mathcal{I}} \hat{\pi}_i \sum_{z\in\mathcal{S}_i} \hat{p}_{i,z} \log\left(\mathbb{S}_z(\boldsymbol{W}\bar{\boldsymbol{h}}_i)\right) .$$

At this point note that this is of identical form to (2). Consequently, the definitions (e.g., NTP-separability, NTP-margin) and results derived in the main body for sequences of fixed length are applicable to the AR setting, extending mutatis mutandis.

*Remark* D.1 (The role of sequence length.). Despite the above reduction of the AR setting to the fixed-length setting, it is crucial to recognize that sequence length remains a significant factor in the AR model. Specifically, it influences the formulation through support sets and their associated probabilities. As sequences extend in length, their corresponding support sets generally become sparser, indicative of less ambiguity in predicting the next token. This dynamic is captured by Shannon's inequality,

$$\mathcal{H}_t \geq \mathcal{H}_{t+1}, \text{ where } \mathcal{H}_t = -\sum_{j\in[m_t]} \sum_{z\in\mathcal{S}^{\ell}_{t,j}} \pi_{t,j}\hat{p}_{t,j,z} \log(\hat{p}_{t,j,z}),$$

reflecting the incremental reduction in entropy as sequence length increases.

## E Notations

Throughout, lowercase and uppercase bold letters (e.g., $\boldsymbol{a}$ and $\boldsymbol{A}$) represent vectors and matrices, respectively. $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote Euclidean inner product and norm, respectively. For matrix $\boldsymbol{A}$, we denote its pseudoinverse as $\boldsymbol{A}^{\dagger}$. All logarithms are natural logarithms (base $e$). We denote $\boldsymbol{e}_v$ the $v$-th standard basis vector in $\mathbb{R}^V$. $\Delta^{V-1}$ denotes the $V$-dimensional unit simplex and $\mathbb{S}() : \mathbb{R}^V \to \Delta^{V-1}$ the softmax map:

$$\mathbb{S}(\boldsymbol{a}) = [\mathbb{S}_1(\boldsymbol{a}), \ldots, \mathbb{S}_V(\boldsymbol{a})]^{\top}, \qquad \text{with } \mathbb{S}_v(\boldsymbol{a}) = \frac{e^{\boldsymbol{e}_v^{\top}\boldsymbol{a}}}{\sum_{v'\in[V]} e^{\boldsymbol{e}_{v'}^{\top}\boldsymbol{a}}} .$$

As explained in Section 2 we represent a training set as

$$\mathcal{T}_m := \{(\bar{\boldsymbol{h}}_j, \hat{\pi}_j, \hat{p}_{j,z\in\mathcal{V}})\}_{j\in[m]} .$$

We assume that embeddings are bounded and denote

$$M := \sqrt{2} \max_{j\in[m]} \|\bar{\boldsymbol{h}}_j\| .$$

Given $\mathcal{T}_m$, let

$$\mathcal{F} = \mathrm{span}\left(\left\{(\boldsymbol{e}_z - \boldsymbol{e}_{z'})\bar{\boldsymbol{h}}_j^{\top} \,:\, z \neq z' \in \mathcal{S}_j, j \in [m]\right\}\right)$$

a subspace of $V \times d$ matrices and $\mathcal{F}^{\perp}$ its orthogonal complement. Denote $\mathcal{P}_{\mathcal{F}}, \mathcal{P}_{\perp}$ the orthogonal projections onto $\mathcal{F}$ and $\mathcal{F}^{\perp}$, respectively. For convenience, for $\boldsymbol{W} \in \mathbb{R}^{V\times d}$, we denote

$$\boldsymbol{W}_{\mathcal{F}} := \mathcal{P}_{\mathcal{F}}(\boldsymbol{W}) \qquad \text{and} \qquad \boldsymbol{W}_{\perp} = \mathcal{P}_{\perp}(\boldsymbol{W}) .$$

Define

$$\mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}) = \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(1 + \sum_{z\neq z} e^{-(\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \boldsymbol{W}\bar{\boldsymbol{h}}_j}\right). \tag{7}$$

Clearly, for all $\boldsymbol{W} \in \mathbb{R}^{V\times d}$, it holds $\mathrm{CE}(\boldsymbol{W}) \geq \mathrm{CE}_{\mathcal{F}}(\boldsymbol{W})$. Note also that for all $\boldsymbol{W} \in \mathcal{F}$ and for all $\boldsymbol{W}^{\mathrm{d}} \in \mathcal{F}^{\perp}$ that satisfy Eq. (4a), it holds $\mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}) = \lim_{R\to\infty} \mathrm{CE}(\boldsymbol{W} + R\boldsymbol{W}^{\mathrm{d}})$. Thus, under NTP compatibility and NTP separability,

$$\inf_{\boldsymbol{W}\in\mathcal{F}} \mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}) = \inf_{\boldsymbol{W}} \mathrm{CE}(\boldsymbol{W}) = \mathcal{H}. \tag{8}$$

# F Proofs

## F.1 Gradient Descent

Throughout we assume GD is ran with step-size $\eta \leq 1/(2L)$ where $L$ is the smoothness of CE loss. This condition is not explicitly mentioned thereafter.

### F.1.1 Auxiliary Lemmata

The following result follows from standard optimization analysis for smooth convex functions specialized to functions that do not attain their infimum. The version presented here is adopted from Lemma 2 (Ji et al., 2020).

**Lemma F.1.** *It holds*

$$\lim_{k\to\infty} \mathrm{CE}(\boldsymbol{W}_k) = \inf_{\boldsymbol{W}} \mathrm{CE}(\boldsymbol{W})$$

*and also* $\lim_{k\to\infty} \|\boldsymbol{W}_k\| = \infty$.

In the lemma below, we collect some useful and simple-to-show properties of the GD and regularization paths. Analogous results, for the different setting of one-hot binary classification over general non-separable data have been established in (Ji and Telgarsky, 2018).

**Lemma F.2.** *Suppose conditions (4) hold for some $\boldsymbol{W}^{\mathrm{d}}$. Also, that there exists $\boldsymbol{W}^{\mathrm{p}} = \boldsymbol{W}^\star \in \mathcal{F}$ satisfying condition (3). The following hold:*

1. $\mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}^\star) = \inf_{\boldsymbol{W}\in\mathcal{F}} \mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}) = \mathcal{H}$,
2. $\boldsymbol{W}^\star$ *is the* unique *minimizer of* $\mathrm{CE}_{\mathcal{F}}$ *on the subspace* $\mathcal{F}$,
3. $\lim_{k\to\infty} \mathcal{P}_{\mathcal{F}}(\boldsymbol{W}_k) = \boldsymbol{W}^\star$, *where* $\boldsymbol{W}_k$ *are GD iterates,*
4. $\lim_{k\to\infty} \|\mathcal{P}_{\perp}(\boldsymbol{W}_k)\| = \infty$,
5. $\lim_{B\to\infty} \mathcal{P}_{\mathcal{F}}(\widehat{\boldsymbol{W}}_B) = \boldsymbol{W}^\star$, *where* $\widehat{\boldsymbol{W}}_B$ *is the reguarlized solution (6),*
6. $\lim_{B\to\infty} \|\mathcal{P}_{\perp}(\widehat{\boldsymbol{W}}_B)\| = \infty$.

*Proof.* It is easy to check by direct substitution of $\boldsymbol{W}^\star$ in (7) and use of (3) that $\mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}^\star) = \mathcal{H}$. This and (8) show the first claim.

The first claim shows $\boldsymbol{W}^\star$ is a minimizer. Suppose for the sake of contradiction there is a different minimizer $\boldsymbol{W}^\star \neq \boldsymbol{W}_1 \in \mathcal{F}$. Then, since $\mathrm{CE}_{\mathcal{F}}(\boldsymbol{W}_1) = \mathcal{H}$, it also holds for $\boldsymbol{W}_R := \boldsymbol{W}_1 + R\boldsymbol{W}^{\mathrm{d}}$ that $\lim_{R\to\infty} \mathrm{CE}(\boldsymbol{W}_R) = \mathcal{H}$. In turn, this implies for all $j \in [m]$:

$$\lim_{R\to\infty} \mathbb{S}_z(\boldsymbol{W}_R\bar{\boldsymbol{h}}_j) = \hat{p}_{j,z}, \forall z \in \mathcal{S}_j\,, \qquad \text{and} \qquad \lim_{R\to\infty} \mathbb{S}_v(\boldsymbol{W}_R\bar{\boldsymbol{h}}_j) = 0, \forall v \notin \mathcal{S}_j\,.$$

The first condition gives then that $\boldsymbol{W}_1$ must satisfy (3). Since $\boldsymbol{W}^\star$ also satisfies these equations, denoting $\boldsymbol{W}_\Delta = \boldsymbol{W}^\star - \boldsymbol{W}_1 \neq 0$, it holds:

$$\langle \boldsymbol{W}_\Delta, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \bar{\boldsymbol{h}}_j)\rangle = 0,\ \forall j \in [m], z \neq z' \in \mathcal{S}_j.$$

But $\boldsymbol{W}_\Delta \in \mathcal{F}$, so this forms a contradiction. Hence, $\boldsymbol{W}^\star$ is unique solution in $\mathcal{F}$ of (3) and unique minimizer of $\mathrm{CE}_\mathcal{F}$ on the subspace $\mathcal{F}$.

The proof of the third claim follows the same way as the proof of part (1) of Thm. 15 of (Ji et al., 2020). For completeness: It follows by the lemma's assumptions and Lemma F.1 that $\lim_{k\to\infty} \mathrm{CE}(\boldsymbol{W}_k) = \mathcal{H}$. Combining with the first claim of the lemma yields $\lim_{k\to\infty} \mathrm{CE}(\boldsymbol{W}_k) = \mathrm{CE}_\mathcal{F}(\boldsymbol{W}^\star)$. Since $\mathrm{CE}_\mathcal{F}(\boldsymbol{W}_k) \le \mathrm{CE}(\boldsymbol{W}_k)$, this finally gives

$$\lim_{k\to\infty} \mathrm{CE}_\mathcal{F}(\boldsymbol{W}_k) = \lim_{k\to\infty} \mathrm{CE}_\mathcal{F}\left(\mathcal{P}_\mathcal{F}(\boldsymbol{W}_k)\right) = \mathrm{CE}_\mathcal{F}(\boldsymbol{W}^\star).$$

Since $\boldsymbol{W}^\star$ is unique by the second claim, the desired then follows.

For the fourth claim, recall from Lemma F.1 that $\lim_{k\to\infty} \|\boldsymbol{W}_k\| = \infty$. From the previous claim, we also have $\lim_{k\to\infty} \|\mathcal{P}_\mathcal{F}(\boldsymbol{W}_k)\| < C$ for some constant $C > \|\boldsymbol{W}^\star\|$. Thus, the desired follows by applying the fact that $\|\boldsymbol{W}_k\| = \|\mathcal{P}_\mathcal{F}(\boldsymbol{W}_k)\| + \|\mathcal{P}_\perp(\boldsymbol{W}_k)\|$.

The proof of the last two claim is exactly same as that of the third and fourth claim. Only now use the facts that $\lim_{B\to\infty} \mathrm{CE}(\boldsymbol{W}_B) = \mathcal{H}$ and $\lim_{B\to\infty} \|\boldsymbol{W}_B\| = \infty$ (see proof of Theorem A.3). □

### F.1.2 Key Lemma

**Lemma F.3.** *Let $\boldsymbol{W}_k$ denote the GD iterate at iteration $k$. Recall the decomposition $\boldsymbol{W}_k = \mathcal{P}_\mathcal{F}(\boldsymbol{W}_k) + \mathcal{P}_\perp(\boldsymbol{W}_k) = \boldsymbol{W}_{k,\mathcal{F}} + \boldsymbol{W}_{k,\perp}$. Fix any $\alpha \in (0,1)$. There exists large enough $R = R(\alpha)$ and $k_0 = k_0(R)$ such that for any $k \ge k_0$, it holds that $\|\boldsymbol{W}_{k,\perp}\| \ge R$ and*

$$\mathrm{CE}\left(\boldsymbol{W}_{k,\mathcal{F}} + (1+\alpha)\|\boldsymbol{W}_{k,\perp}\|\overline{\boldsymbol{W}^{\mathrm{mm}}}\right) \le \mathrm{CE}(\boldsymbol{W}_k)\,. \tag{9}$$

*Proof.* We drop the subscript $k$ to lighten notation.

First, note by Lemma F.2.D that, for arbitrary $R$, we can pick $k_1 = k_1(R)$ such that for all $k \ge k_1$: $\|\boldsymbol{W}_\perp\| \ge R$.

Thus next, we will prove the main claim, i.e. for large enough $\|\boldsymbol{W}_\perp\|$ inequality (9) holds. Denote $R' = \frac{\|\boldsymbol{W}_\perp\|}{\|\boldsymbol{W}^{\mathrm{mm}}\|}$. Substituting in CE expression (2), and using the fact that $\boldsymbol{W}^{\mathrm{mm}} \in \mathcal{F}^\perp$ by (4a) yield:

$$\begin{aligned}
&\mathrm{CE}\left(\boldsymbol{W}_\mathcal{F} + (1+\alpha)R'\boldsymbol{W}^{\mathrm{mm}}\right)\\
&= \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\sum_{z'\in\mathcal{S}_j} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_{z'})^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-(1+\alpha)R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{mm}}\bar{\boldsymbol{h}}_j}\right).\\
&= \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\sum_{v\in\mathcal{V}} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-(1+\alpha)R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{mm}}\bar{\boldsymbol{h}}_j}\right).
\end{aligned} \tag{10}$$

Moreover, decomposing $\boldsymbol{W} = \boldsymbol{W}_\mathcal{F} + \boldsymbol{W}_\perp$, and defining

$$\widetilde{\boldsymbol{W}}_\perp := \frac{\|\boldsymbol{W}^{\mathrm{mm}}\|}{\|\boldsymbol{W}_\perp\|}\boldsymbol{W}_\perp = \frac{1}{R}\boldsymbol{W}_\perp\,,$$

we have

$$\begin{aligned}
\mathrm{CE}\left(\boldsymbol{W}\right) &= \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\sum_{z'\in\mathcal{S}_j} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_{z'})^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp\bar{\boldsymbol{h}}_j}\right)\\
&= \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\sum_{v\in\mathcal{V}} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_\mathcal{F}\bar{\boldsymbol{h}}_j} + \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp\bar{\boldsymbol{h}}_j}\right),
\end{aligned} \tag{11}$$

where we used that, by definition, $\boldsymbol{W}_\perp \in \mathcal{F}^\perp$. Thus, our goal becomes showing (10) ≤ (11), for large enough $R$. To do this, we consider two cases as follows below.

For the remaining of the proof recall $M := \max_{j\in[m]} \sqrt{2}\|\bar{\boldsymbol{h}}_j\|$ and use the logits shorthand:

$$\widetilde{\ell}_{j,v} = \boldsymbol{e}_v^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_j \qquad \text{and} \qquad \ell^{\mathrm{mm}}_{j,v} = \boldsymbol{e}_v^\top \boldsymbol{W}^{\mathrm{mm}} \bar{\boldsymbol{h}}_j\,.$$

*Case 1: $\boldsymbol{W}_\perp$ is well aligned with $\boldsymbol{W}^{\mathrm{mm}}$.* Suppose

$$\|\boldsymbol{W}^{\mathrm{mm}} - \widetilde{\boldsymbol{W}}_\perp\| \le \epsilon := \frac{\alpha}{M}. \tag{12}$$

Using this, linearity of logits, and Cauchy-Schwartz, yields

$$\widetilde{\ell}_{j,z} - \widetilde{\ell}_{j,v} \le \ell^{\mathrm{mm}}_{j,z} - \ell^{\mathrm{mm}}_{j,v} + \epsilon M, \quad \forall j \in [m], z \in \mathcal{S}_j, v \notin \mathcal{S}_j\,.$$

Thus,

$$\sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_j} \ge e^{-\epsilon M R'} \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{mm}} \bar{\boldsymbol{h}}_j} = e^{-\alpha R'} \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{mm}} \bar{\boldsymbol{h}}_j}$$

Also recall by feasibility of $\boldsymbol{W}^{\mathrm{mm}}$ that

$$\ell^{\mathrm{mm}}_{j,z} - \ell^{\mathrm{mm}}_{j,v} \ge 1, \forall j \in [m], z \in \mathcal{S}_j, v \notin \mathcal{S}_j\,. \tag{13}$$

Thus,

$$\sum_{v\notin\mathcal{S}_j} e^{-(1+\alpha)R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_j} \le e^{-\alpha R'} \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}^{\mathrm{mm}} \bar{\boldsymbol{h}}_j}$$

Comparing the above two displays yields

$$\sum_{v\notin\mathcal{S}_j} e^{-(1+\alpha)R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_j} \le \sum_{v\notin\mathcal{S}_j} e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_j},$$

which implies the desired (10)≤(11) for any value of $R'$ (eqv. $\|\boldsymbol{W}_\perp\|$).

*Case 2: No alignment.* Suppose now that (12) does not hold. Note that $\|\widetilde{\boldsymbol{W}}_\perp\| = \|\boldsymbol{W}^{\mathrm{mm}}\|$ and since (NTP-SVM) has a unique solution it must be that $\widetilde{\boldsymbol{W}}_\perp$ is not feasible. But $\widetilde{\boldsymbol{W}}_\perp \in \mathcal{F}_\perp$, thus it satisfies the equality constraints. This then means that there exist $\delta := \delta(\epsilon)$ and $j_\star \in [m], v_\star \notin \mathcal{S}_{j_\star}$ such that

$$\widetilde{\ell}_{j_\star,z} - \widetilde{\ell}_{j_\star,v_\star} \le 1 - \delta\,, \quad \forall z \in \mathcal{S}_{j_\star}. \tag{14}$$

(Note the above holds for all $z \in \mathcal{S}_{j_\star}$ because $\widetilde{\ell}_{j_\star,z} = \widetilde{\ell}_{j_\star,z'}$ since $\widetilde{\boldsymbol{W}}_\perp \in \mathcal{F}_\perp$.)

To continue, we introduce the shorthand notation

$$A_{j,z} := A_{j,z}(\boldsymbol{W}) = \sum_{v\in\mathcal{V}} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_{\mathcal{F}} \bar{\boldsymbol{h}}_j}$$

as well as

$$A_{\min} := \min_{j\in[m], z\in\mathcal{S}_j} A_{j,z}, \qquad \text{and} \qquad A_{\max} := \max_{j\in[m], z\in\mathcal{S}_j} A_{j,z}\,.$$

Using (14) we may lower bound (11) as follows:

$$\begin{aligned}
\mathrm{CE}(\boldsymbol{W}) - \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\sum_{v\in\mathcal{V}} e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top \boldsymbol{W}_{\mathcal{F}} \bar{\boldsymbol{h}}_j}\right) &\ge \hat{\pi}_{j_\star} \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(1 + \frac{e^{-R'(\boldsymbol{e}_z-\boldsymbol{e}_{v_\star})^\top \widetilde{\boldsymbol{W}}_\perp \bar{\boldsymbol{h}}_{j_\star}}}{A_{j_\star,z}}\right) \\
&\ge \hat{\pi}_{j_\star} \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(1 + \frac{e^{-R'(1-\delta)}}{A_{\max}}\right) \\
&\ge \frac{e^{-R'(1-\delta)}}{n(A_{\max}+1)}\,,
\end{aligned} \tag{15}$$

where in the last line we used $\hat{\pi}_j \ge 1/n, \forall j \in [m]$ as well as $\log(1+x) \ge \frac{x}{1+x}, x > 0$.

On the other hand, using property (13) for max-margin logits, we can upper bound (10) as follows:

$$\mathrm{CE}\left(\boldsymbol{W}_{\mathcal{F}}+(1+\alpha)R'\boldsymbol{W}^{\mathrm{mm}}\right)-\sum_{j\in[m]}\hat{\pi}_j\sum_{z\in\mathcal{S}_j}\hat{p}_{j,z}\log\left(\sum_{v\in\mathcal{V}}e^{-(\boldsymbol{e}_z-\boldsymbol{e}_v)^\top\boldsymbol{W}_{\mathcal{F}}\bar{\boldsymbol{h}}_j}\right)\le\log\left(1+\frac{Ve^{-R'(1+\alpha)}}{A_{\min}}\right)$$
$$\le\frac{Ve^{-R'(1+\alpha)}}{A_{\min}},\tag{16}$$

where in the last line we used $\log(1+x)\le x, x>0$.

In view of the two last displays, it suffices that

$$V\frac{e^{-R'(1+\alpha)}}{A_{\min}}\le\frac{e^{-R'(1-\delta)}}{n(A_{\max}+1)}\iff R'\ge\frac{1}{\delta+\alpha}\log\left(\frac{nV(A_{\max}+1)}{A_{\min}}\right).$$

All it remains is obtaining bounds for $A_{\min}, A_{\max}$ specifically showing that they do not depend on $R$. By Cauchy-Schwartz:

$$Ve^{-M\|\boldsymbol{W}_{\mathcal{F}}\|}\le\boldsymbol{A}_{\min}\le\boldsymbol{A}_{\max}\le Ve^{M\|\boldsymbol{W}_{\mathcal{F}}\|}$$

Further recall by Lemma F.2.C that if $k$ is large enough then

$$\|\boldsymbol{W}_{\mathcal{F}}-\boldsymbol{W}^\star\|\le\|\boldsymbol{W}^\star\|\implies\|\boldsymbol{W}_{\mathcal{F}}\|\le2\|\boldsymbol{W}^\star\|.\tag{17}$$

Thus, there exists $k_\star=k_\star(\|\boldsymbol{W}_\star\|)$ such that for all $k\ge k_\star$:

$$Ve^{-2M\|\boldsymbol{W}_\star\|}\le\boldsymbol{A}_{\min}\le\boldsymbol{A}_{\max}\le Ve^{2M\|\boldsymbol{W}_\star\|}.$$

Hence, the desired (16)≤(15) holds provided

$$\|\boldsymbol{W}_\perp\|\ge\frac{\|\boldsymbol{W}^{\mathrm{mm}}\|}{\alpha}\log\left(2nVe^{4\|\boldsymbol{W}^\star\|}\right).\tag{18}$$

Set $R=R(\alpha)=\{\text{RHS of (18)}\}$ and $k_0(R):=\max\{k_1(R),k_\star\}$. We have shown this guarantees for all $k\ge k_0$: $\|\boldsymbol{W}_\perp\|\ge R$ and by choice of $R$ also (16)≤(15). This in turn implies (10)≤(11), as desired to complete the proof. □

### F.1.3 Proof of Theorem 4.1

For the subspace component, see Lemma F.2.C. For the directional convergence, the key ingredient of the proof is Lemma F.3. After that, the proof follows identically to Thm. 15(2) (Ji et al., 2020). We include the details for completeness, but there are no novel aspects in the rest of this section.

Let any $\epsilon\in(0,1)$ and choose $\alpha=\epsilon/(1-\epsilon)$. By Lemma F.3, there exists $k_0$ such that for any $k\ge k_0$, we have

$$\|\boldsymbol{W}_{k,\perp}\|\ge\max\{R(\alpha),1/2\}$$

and

$$\langle\nabla\mathrm{CE}(\boldsymbol{W}_k),\boldsymbol{W}_{k,\perp}-(1+\alpha)\|\boldsymbol{W}_{k,\perp}\|\overline{\boldsymbol{W}^{\mathrm{mm}}}\rangle=\langle\nabla\mathrm{CE}(\boldsymbol{W}_k),\boldsymbol{W}_k-(\boldsymbol{W}_{k,\mathcal{F}}+(1+\alpha)\|\boldsymbol{W}_{k,\perp}\|\overline{\boldsymbol{W}^{\mathrm{mm}}})\rangle$$
$$\ge\mathrm{CE}(\boldsymbol{W}_k)-\mathrm{CE}(\boldsymbol{W}_{k,\mathcal{F}}+(1+\alpha)\|\boldsymbol{W}_{k,\perp}\|\overline{\boldsymbol{W}^{\mathrm{mm}}})\ge0,$$

where we also used convexity of the loss.

Consequently,

$$\begin{aligned}
\langle\boldsymbol{W}_{k+1}-\boldsymbol{W}_k,\overline{\boldsymbol{W}^{\mathrm{mm}}}\rangle&=\langle-\eta\nabla\mathrm{CE}(\boldsymbol{W}_k),\overline{\boldsymbol{W}^{\mathrm{mm}}}\rangle\\
&\ge(1-\epsilon)\langle-\eta\nabla\mathrm{CE}(\boldsymbol{W}_k),\overline{\boldsymbol{W}_{k,\perp}}\rangle\\
&\ge(1-\epsilon)\langle\boldsymbol{W}_{k+1,\perp}-\boldsymbol{W}_{k,\perp},\overline{\boldsymbol{W}_{k,\perp}}\rangle\\
&\ge(1-\epsilon)\langle\boldsymbol{W}_{k+1,\perp}-\boldsymbol{W}_{k,\perp},\overline{\boldsymbol{W}_{k,\perp}}\rangle\\
&=\frac{(1-\epsilon)}{2\|\boldsymbol{W}_{k,\perp}\|}\left(\|\boldsymbol{W}_{k+1,\perp}\|^2-\|\boldsymbol{W}_{k,\perp}\|^2-\|\boldsymbol{W}_{k+1,\perp}-\boldsymbol{W}_{k,\perp}\|^2\right)\\
&\ge(1-\epsilon)\left(\|\boldsymbol{W}_{k+1,\perp}\|-\|\boldsymbol{W}_{k,\perp}\|-2\eta(\mathrm{CE}(\boldsymbol{W}_{k,\perp})-\mathrm{CE}(\boldsymbol{W}_{k+1,\perp}))\right),
\end{aligned}$$

where the last step used $\|\boldsymbol{W}_{k,\perp}\| \geq 1/2$, the fact that $x^2 - y^2 \geq 2y(x-y), \forall x, y$ and smoothness of the CE loss.

Telescoping the above expression and rearranging yields

$$\begin{aligned}\langle \overline{\boldsymbol{W}}_k, \overline{\boldsymbol{W}^{\mathrm{mm}}} \rangle &\geq (1-\epsilon)\frac{\|\boldsymbol{W}_{k,\perp}\|}{\|\boldsymbol{W}_k\|} - \frac{\langle \boldsymbol{W}_{k_0}, \overline{\boldsymbol{W}^{\mathrm{mm}}} \rangle - (1-\epsilon)\|\boldsymbol{w}_{k_0,\perp}\| - \eta\,\mathrm{CE}(\boldsymbol{W}_{k_0})}{\|\boldsymbol{W}_k\|} \\ &\geq (1-\epsilon) - \frac{\|\boldsymbol{W}_{k,\mathcal{F}}\|_2 + \langle \boldsymbol{W}_{k_0}, \overline{\boldsymbol{W}^{\mathrm{mm}}} \rangle - (1-\epsilon)\|\boldsymbol{w}_{k_0,\perp}\| - \eta\,\mathrm{CE}(\boldsymbol{W}_{k_0})}{\|\boldsymbol{W}_k\|}\end{aligned}$$

Now recall from Lemma F.2 that $\lim_{k\to\infty}\|\boldsymbol{W}_k\| = \infty$ and $\lim_{k\to\infty}\|\boldsymbol{W}_{k,\mathcal{F}}\| = \|\boldsymbol{W}^\star\|$. Thus, $\liminf_{k\to\infty}\langle \overline{\boldsymbol{W}}_k, \overline{\boldsymbol{W}^{\mathrm{mm}}} \rangle \geq 1-\epsilon$. Since $\epsilon$ is arbitrary, the desired follows.

### F.2 Regularization Path

We provide a detailed proof of Theorem A.3 filling in missing details from the proof sketch in the main paper.

#### F.2.1 Proof of Theorem A.3

First, we show that $\widehat{\boldsymbol{W}}_B$ is on the boundary, i.e. $\|\widehat{\boldsymbol{W}}_B\| = B$. Suppose not, then $\langle \nabla\,\mathrm{CE}(\widehat{\boldsymbol{W}}_B), \boldsymbol{U}\rangle = 0$ for all $\boldsymbol{U} \in \mathbb{R}^{V\times d}$. Using the CE expression in (2) and a few algebraic manipulations, yields

$$\langle -\nabla\,\mathrm{CE}(\widehat{\boldsymbol{W}}_B), \boldsymbol{U}\rangle = \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z}\Big( \sum_{\substack{z'\in\mathcal{S}_j \\ z'\neq z}} s_{j,z'}\,(\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \boldsymbol{U}\bar{\boldsymbol{h}}_j + \sum_{v\notin\mathcal{S}_j} s_{j,v}\,(\boldsymbol{e}_z - \boldsymbol{e}_v)^\top \boldsymbol{U}\bar{\boldsymbol{h}}_j \Big), \tag{19}$$

where we denote the output probabilities at $\widehat{\boldsymbol{W}}_B$ as $s_{j,v} := \mathbb{S}_v(\widehat{\boldsymbol{W}}_B\bar{\boldsymbol{h}}_j), v \in \mathcal{V}, j \in [m]$. Choose $\boldsymbol{U} = \boldsymbol{W}^{\mathrm{mm}}$ in (19). Then, the first term in the parenthesis in (19) is zero by (4a), while the second term is strictly positive by (4b) and strict positivity of softmax entries, leading to contradiction.

Now, consider point $\boldsymbol{W}_B^\star = \boldsymbol{W}^\star + R(B)\cdot\boldsymbol{W}^{\mathrm{mm}}$, where, $\boldsymbol{W}^\star \in \mathcal{T}$ satisfies (3), and $R = R(B)$ is chosen such that $\|\boldsymbol{W}_B^\star\| = B$. Concretely, for $B > \|\boldsymbol{W}^\star\|$, set

$$R = \frac{1}{\|\boldsymbol{W}^{\mathrm{mm}}\|}\sqrt{B^2 - \|\boldsymbol{W}^\star\|^2}.$$

Note also that $R/B \to 1/\|\boldsymbol{W}^{\mathrm{mm}}\|$ as $B\to\infty$. We will show that $\boldsymbol{W}_B^\star$ attains a small CE loss as $B$ (hence, $R$) grows. To do this, denote for convenience the logits for all $v \in \mathcal{V}, j \in [m]$ :

$$\ell_{j,v}^\star := \boldsymbol{e}_v^\top \boldsymbol{W}^\star \bar{\boldsymbol{h}}_j \quad \text{and} \quad \ell_{j,v}^{\mathrm{mm}} := \boldsymbol{e}_v^\top \boldsymbol{W}^{\mathrm{mm}} \bar{\boldsymbol{h}}_j\,,$$

and note that $\boldsymbol{e}_v^\top \boldsymbol{W}_B^\star \bar{\boldsymbol{h}}_j = \ell_{j,v}^\star + R\,\ell_{j,v}^{\mathrm{mm}}$. By using (3) and (4a):

$$\sum_{z'\in\mathcal{S}_j} e^{-(\ell_{j,z}^\star + R\ell_{j,z}^{\mathrm{mm}} - \ell_{j,z'}^\star - R\ell_{j,z'}^{\mathrm{mm}})} = \frac{1}{\hat{p}_j}.$$

Moreover, using (4b)

$$\sum_{v\notin\mathcal{S}_j} e^{-(\ell_{j,z}^\star + R\ell_{j,z}^{\mathrm{mm}} - \ell_{j,v}^\star - R\ell_{j,v}^{\mathrm{mm}})} \leq e^{-R}\sum_{v\notin\mathcal{S}_j} e^{-(\ell_{j,z}^\star - \ell_{j,v}^\star)} \leq C\,e^{-R},$$

where we define constant (independent of $R$) $C := Ve^{\|\boldsymbol{W}^\star\|M}$, for $M := \sqrt{2}\cdot\max_{j/\in[m]}\|\bar{\boldsymbol{h}}_j\|$.

Combining the above displays and using in Eq. (2), yields

$$\begin{aligned}\mathrm{CE}(\boldsymbol{W}_B^\star) &\leq \sum_{j\in[m]}\hat{\pi}_j\sum_{z\in\mathcal{S}_j}\hat{p}_{j,z}\log\Big(\frac{1}{\hat{p}_{j,z}} + C\,e^{-R}\Big) \leq \sum_{j\in[m]}\hat{\pi}_j\sum_{z\in\mathcal{S}_j}\hat{p}_{j,z}\Big(\log\big(\frac{1}{\hat{p}_{j,z}}\big) + \hat{p}_{j,z}C\,e^{-R}\Big) \\ &\leq \mathcal{H} + C\,e^{-R}\,,\end{aligned} \tag{20}$$

where, the second line uses $\log(1+x) \le x, x > 0$, and the third line uses $\hat{\pi}_j, \hat{p}_{j,z}$ are probabilities.

Next, towards arriving at a contradiction, we will show that if $\widehat{\boldsymbol{W}}_B$ is not in the direction of $\boldsymbol{W}^{\mathrm{mm}}$, then it incurs a loss that is larger than $\mathrm{CE}(\boldsymbol{W}_B^\star)$. Concretely, assuming the statement of the theorem is not true, we we will upper bound

$$\mathrm{CE}(\widehat{\boldsymbol{W}}_B) - \mathcal{H} = \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\Big(\frac{\hat{p}_{j,z}}{\mathbb{S}_z(\widehat{\boldsymbol{W}}_B \bar{\boldsymbol{h}}_j)}\Big). \tag{21}$$

By our assumption, there exists $\epsilon > 0$, such that there exists arbitrarily large $B$ satisfying:

$$\Big\| \frac{\|\boldsymbol{W}^{\mathrm{mm}}\|}{B} \widehat{\boldsymbol{W}}_B - \boldsymbol{W}^{\mathrm{mm}} \Big\| > \epsilon. \tag{22}$$

Define

$$\widehat{\boldsymbol{W}} = \frac{1}{R'(B)} \big(\widehat{\boldsymbol{W}}_B - \boldsymbol{W}^\star\big),$$

where, $R' = R'(B) > 0$ is chosen so that $\|\widehat{\boldsymbol{W}}\| = \|\boldsymbol{W}^{\mathrm{mm}}\|$. Concretely, for large enough $B \ge 2\|\boldsymbol{W}^{\mathrm{mm}}\|$, set

$$R' = \sqrt{\frac{B^2}{\|\boldsymbol{W}^{\mathrm{mm}}\|^2} - 2B\langle \overline{\boldsymbol{W}_B}, \overline{\boldsymbol{W}^{\mathrm{mm}}}\rangle + 1}\,.$$

Note that it holds $\lim_{B\to\infty} R'/B = 1/\|\boldsymbol{W}^{\mathrm{mm}}\|$. Thus, we can always choose $B$ large enough so that Eq. (22) guarantees $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^{\mathrm{mm}}\| \ge \epsilon'$, for some $\epsilon' > 0$. Since $\boldsymbol{W}^{\mathrm{mm}}$ is the unique minimizer of (NTP-SVM) and $\|\widehat{\boldsymbol{W}}\| = \|\boldsymbol{W}^{\mathrm{mm}}\|$, it follows that there exists $\delta \in (0,1)$ and $j \in [m]$ such that at least one of the following is true

(i) $\exists z$ and $z' \ne z \in \mathcal{S}_j$ such that

$$|(\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \widehat{\boldsymbol{W}} \bar{\boldsymbol{h}}_j| \ge \delta\,, \tag{23}$$

(ii) $\exists z \in \mathcal{S}_j, v \notin \mathcal{S}_j$ such that

$$(\boldsymbol{e}_z - \boldsymbol{e}_v)^\top \widehat{\boldsymbol{W}} \bar{\boldsymbol{h}}_j \le 1 - \delta. \tag{24}$$

*Case (i):* Without loss of generality $(\boldsymbol{e}_z - \boldsymbol{e}_{z'})^\top \widehat{\boldsymbol{W}} \bar{\boldsymbol{h}}_j \le -\delta$ (otherwise, flip $z, z'$). Thus, ignoring all but one term in (21) gives

$$\mathrm{CE}(\widehat{\boldsymbol{W}}_B) - \mathcal{H} \ge \hat{\pi}_j \hat{p}_{j,z} \log\Big(\frac{\hat{p}_{j,z}}{\mathbb{S}_z(\widehat{\boldsymbol{W}}_B \bar{\boldsymbol{h}}_j)}\Big) \ge \hat{\pi}_j \hat{p}_{j,z} \log\Big(\hat{p}_{j,z} e^{(\ell_{j,z'} - \ell_{j,z})}\Big), \tag{25}$$

where we use $\ell_{j,v} = \boldsymbol{e}_v^\top \widehat{\boldsymbol{W}}_B \bar{\boldsymbol{h}}_j, v \in \mathcal{V}$ to denote logits of $\widehat{\boldsymbol{W}}_B$. Using (3) and (23), yields

$$\ell_{j,z'} - \ell_{j,z} = (\boldsymbol{e}_{z'} - \boldsymbol{e}_z)^\top \big(R' \widehat{\boldsymbol{W}} + \boldsymbol{W}^\star\big) \bar{\boldsymbol{h}}_j \ge R'\delta + \log\left(\frac{\hat{p}_{j,z'}}{\hat{p}_{j,z}}\right).$$

Put in (21) and using $\hat{p}_{j,z} \ge \hat{\pi}_j \hat{p}_{j,z} \ge 1/n$ shows

$$\mathrm{CE}(\widehat{\boldsymbol{W}}_B) \ge \mathcal{H} + \frac{1}{n} \log\Big(\frac{e^{R'\delta}}{n}\Big)$$

Compare this with (20). For large enough $B$, it is clear that $\hat{\pi}_j \hat{p}_{j,z} \log\Big(\hat{p}_{j,z}\, c\, e^{R'\delta}\Big) > C e^{-R}$. Thus, $\mathrm{CE}(\widehat{\boldsymbol{W}}_B) > \mathrm{CE}(\boldsymbol{W}_B^\star)$, a contradiction.

*Case (ii):* We can assume $\widehat{\boldsymbol{W}} \in \mathcal{T}_\perp$, since otherwise we are in Case (i). Now, again ignoring all but the $(j,z)$ term in the CE loss for which (24) holds for some $v \notin \mathcal{S}_j$, we find

$$\mathrm{CE}(\widehat{\boldsymbol{W}}_B) - \mathcal{H} \ge \hat{\pi}_j \hat{p}_{j,z} \log\Big(\hat{p}_{j,z}\big(\sum_{z'\in\mathcal{S}_j} e^{(\ell_{j,z'} - \ell_{j,z})} + e^{(\ell_{j,v} - \ell_{j,z})}\big)\Big).$$

Using $\mathcal{P}_{\mathcal{T}}(\widehat{\boldsymbol{W}}_B) = \boldsymbol{W}^\star$ yields

$$\sum_{z' \in \mathcal{S}_j} e^{(\ell_{j,z'} - \ell_{j,z})} = \sum_{z' \in \mathcal{S}_j} \frac{\hat{p}_{j,z'}}{\hat{p}_{j,z}} = \frac{1}{\hat{p}_{j,z}} .$$

Moreover, by (24):

$$e^{\ell_{j,v} - \ell_{j,z}} \geq e^{-R'(1-\delta)} \, e^{\ell^\star_{j,v} - \ell^\star_{j,z}} \geq c' e^{-R'(1-\delta)},$$

for constant (independent of $B$) $c' := e^{-\|\boldsymbol{W}^\star\| M}$. Putting the above together yield:

$$\mathrm{CE}(\widehat{\boldsymbol{W}}_B) - \mathcal{H} \geq \hat{\pi}_j \hat{p}_{j,z} \log\left(1 + \hat{p}_{j,z} c' e^{-R'(1-\delta)}\right) \geq \frac{c' e^{-R'(1-\delta)}}{2n^2} .$$

where the second inequality uses $\log(1+x) \geq \frac{x}{1+x}, x > 0$.

Compare this with (20). For large enough $B$, (recall $R, R'$ grow at the same rate) it holds $\frac{c'}{2n^2} e^{-R'(1-\delta)} > C e^{-R}$. Thus, $\mathrm{CE}(\widehat{\boldsymbol{W}}_B) > \mathrm{CE}(\boldsymbol{W}^\star_B)$, a contradiction.

In either case, we arrive at a contradiction, which completes the proof.